An initial investigation in the diagnosis of Alzheimer's disease using various classification techniques

S. R. Bhagya Shree
Research scholar
PET research Center
PES College of Engineering
Mandya, India.
srbhagyashree@yahoo.co.in

Dr. H. S. Sheshadri
Dean, PET research center
Prof. Department of E & C
PES College of Engineering
Mandya, India
hssheshadri@gmail.com

Abstract: ***Now a day's most of the people suffer from brain related neurodegenerative disorders. These disorders lead to various diseases. Dementia is one such disease. Dementia is a general term for a decline in mental ability severe enough to interfere with daily life. Alzheimer's disease is the most common type of dementia. Alzheimer's disease is one of the types of the dementia which accounts to 60-80% of mental disorders [1]. Diagnosis of the disease at the earlier stage is a crucial task. Diagnosis of the disease at the early stage will enable the diseased to have quality life. Authors*** **have** ***collected data from various neuropsychologists which consist of 250 patient's records. In this paper the authors focus on diagnosis of disease using various machine learning techniques of data mining. Authors have compared various classification techniques such as Naive Bayes, Decision tree algorithm J48, Random forest, JRIP and suggest Naïve bayes as the best technique.***

Keywords – CoG, Naive Bayes, Decision tree algorithm J48, Random forest, JRIP, WEKA

I. INTRODUCTION

There are around 44 million people suffering from Dementia [2]. Dementia originated from Latin word where 'de' means 'apart' and 'mentis' means 'mind'. The dementia is classified into various types namely Alzheimer's disease, Vascular dementia, Dementia with Lewy bodies, Front temporal lobar degeneration, Mixed dementia, Parkinson's disease, Creutzfeldt- Jakob disease and Normal pressure hydrocephalus[3]. Alzheimer's disease is one of the types of Dementia. This disease accounts for 60-80% of dementia. There are around 38 million people suffering from Alzheimer's disease [4]. If the disease is not diagnosed at the initial stage the severity of the disease increases. Alzheimer's disease is officially listed as the sixth-leading cause of death in the United States. It is the fifth-leading cause of death for those aged 65 and above. There are various risk factors which contribute to the development of the disease. They are Age, Genetics, Smoking and Alcohol Intake, Cholesterol, Down Syndrome etc. [5]. The symptoms of Alzheimer`s diseases are decision making, poor judgment, misplacing things, impairment of movements, verbal communication, abnormal moods, complete loss of memory.

The diagnosis of AD is done at three different stages namely consulting the General Physician, Undergoing neuro psychological tests and taking MRI scans [6]. Diagnosis of the disease at the early stage will help the patients to have quality life for the rest of their life. So, the authors have focused on diagnosis of the disease for neuro psychological test. The authors have applied various classification techniques of data mining and compared the techniques.

II. LITERATURE SURVEY

Dementia is the disease of the brain, causing loss of cognitive functions like reasoning, memory and other mental abilities due to trauma or normal ageing Alzheimer disease accounts for 60-80% of dementia [7]. If the disease is not diagnosed at the initial stage the severity of the disease increases. The diagnosis is done at three different stages. The first stage is to consult a general physician. The second stage is to undergo various neuro psychological tests and third stage is taking MRI scans [8].This paper focus on diagnosis of AD for a neuro psychological test. There are various neuro psychological tests like MMSE, BDIMC, COG, BOMC, MOCA, AD8 and GP CoG are used. Each of the tests has its own advantage and disadvantage and moreover, the tests are meant for a community of people. Of the all MMSE is very popular. But even that has a disadvantage. The disadvantage of MMSE is it is insensitive to early changes of dementia. This indicates the need of a screening test which may be used to the subjects irrespective of gender, religion, culture and education. The 10/66 Dementia Research Group (10/66) founded in 1998 is a network comprising or more than 100 researchers from many developing countries. 10/66 is committed to good quality research in those regions, where an estimated two-thirds of all those with dementia live. It represents a collaboration of academics, clinicians, and an international non-governmental organization, Alzheimer's disease International (ADI). The 10/66 research group has suggested a battery which fulfills the above requirements. The paper focuses on diagnosis of AD using 10/66 battery by knowledge discovery from data [9]. This 10/66 battery is preferred compared to the most popular MMSE battery as it is applicable to anyone irrespective of gender, religion, culture and education [10]. In this battery a predefined questions will be asked to the subject. Each answer will be evaluated. The score is compared with the score defined by 10/66 research group. Accordingly, the subjects' will be classified as demented or not. Analysis of data and decision making is a crucial step. Many a times the analysis and decision making depends on the mood of the Psychologist. In addition to that, the humane error cannot be avoided. This problem could be overcome by having a machine based analysis. Thus the authors are using data mining approach to discover the

978-1-4799-3975-6/14/$31.00 ©2014 IEEE

knowledge. The knowledge Discovery process is a procedure that comprises of Data Cleaning, Data integration, Data selection, Data Transformation, Data mining, Pattern evaluation, Knowledge presentations [11]. Data mining can be done by using the approaches like statistics, artificial intelligence and machine learning [12]. This paper focuses on various classification techniques used under machine learning and also gives the comparison of the same.

Bhagya Shree S R et.al in their paper have discussed how various researchers have used data miningin the diagnosis of various diseases.[9]. Jyothi soni et.al in their paper used decision tree and Bayesian classification in analyzing the data sets of heart disease patient's [13]. Ruijuan Hu has suggested ID3 algorithm of decision tree in the diagnosis of breast cancer. Decision tree algorithm can classify and predict the various testing data and inspecting data in the medical database to help doctors make an objective and effective in patients with the diagnosis, and help doctors' effective and objective diagnosis [14]. Amir Fallahi et.al have used Bayesian network for detection of breast cancer. They have experimentally proved that Bayesian approach is better than neural approach for breast cancer [15]. Abhishek Taneja in his paper has discussed about using data mining for the prediction of heart disease. To employ the selected classification algorithms four experiments were designed and the experiments were conducted. For all the experiments two settings was done, one containing all the 15 variables and the other containing 8 chosen variables. All the experiments were done on a full training dataset containing all the instances and cross validation was used for randomly sampling the training and test sets. The experimental results have shown that, in general, J48 Decision Tree algorithm outperformed Naïve Bayes classifier and Neural Networks in the domain of predicting heart disease cases [16].Duarte Ferrira et.al have used decision tree classification in the diagnosis of neonantal jaundice as they have an advantage of being more easily interpretable when compared with the closed models, usually called black box models, such as artificial neural networks. Because of this advantage of decision tree it is more easily accepted by medical community [17].

P. Rajeshwari in her paper has used FT tree algorithm in the analysis of liver disorder. FT algorithm is considered as the better performance algorithm. [18].Tarigopppula V.S sriram et.al has used classification algorithms to detect Parkinson's disease [19].

Sandhya Joshi et.al have discussed about the classification of AD, Vascular dementia and Parkinson`s disease using machine learning approach. The entire work demonstrates the effectiveness of considering most influential risk factor for the correct classification of AD, VD and PD. In their paper they have considered the record of 180 subjects. They have used Random forest and multilayer perceptron and achieved an accuracy of 99.33% [20].

Javier Escudero et.al in their paper discussed about detection of Alzheimer's disease using machine learning [21]. Plamena Andreeva et.al have discussed about the various learning models. They have also given the guidelines on the suitable learning model [22].

III. PROBLEM DEFINITION

Data set consists of records of 250 subjects. The records comprises of details of subjects aged from 64-85 years. In this paper the authors focus on classification of subjects as demented or not demented using various machine learning techniques of data mining.

The main objectives are:

- To classify the records as patients of dementia or not by using different Machine Learning techniques namely Naïve Bayes, Decision tree, Random Forest, JRIP.
- Comparison of different parameters.
- Finding out the best classification technique,depending on result.

IV. ARCHITECTURE

Fig.1 Shows the working flow diagram of the paper. The first step is collecting the details of the subjects. This is followed by preprocessing. As the real world data tend to be incomplete, noisy and inconsistent, data preprocessing is an important issue for data mining.

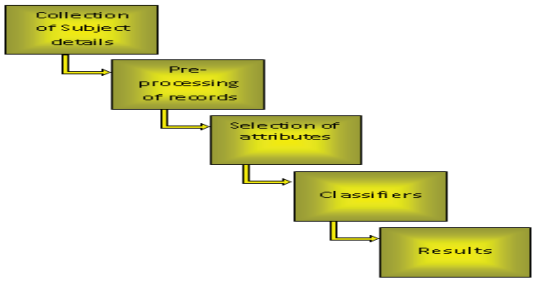

Fig.1: Block diagram of working flow

As the data collected here are primary data, there are no chances of missing of any data. Hence data pre-processing is not required. In the next stage the attributes are selected. Though there are around 24 attributes, the score of certain attributes makes large difference in decision making. The next step is classification. Classification is done to know how exactly the data is being classified. The Classify Tab shows the list of machine learning algorithms. These algorithms in general operate on a classification algorithm and run it multiple times. Algorithm parameters and input data weight may be manipulated to increase the accuracy of the classifier. For analysis purpose WEKA tool is being used. Two learning performance evaluators are included with WEKA. The first is the classifier which simply splits a dataset into training and test data. Second is the cross-validation using folds. Evaluation is usually described by the accuracy. The run information is also displayed, for quick inspection of how well a classifier works.

A. Collection of Datasets.

Datasets consisted of 250 patient records collected from various neuro-psychologists. In that there are four age groups, they are 65-69, 70-75, 76-79 and above 80yrs. The Age group 60-65 consists of 161 records, 70-75 consists of 16 records, 76-79 consists of 22 records and the group of above 80 years of age consists of 51 records.

B. Preprocessing

Pre-processing is a step to check for missing and incorrect values. As there are no chances of missing data, data pre-processing is not performed here.

C. Selection of Attributes

The data set comprises of 24 attributes namely, age, age-group, learning, naming, object description, sentence repetition, word recall, delayed recall, verbal fluency, registration & immediate recall, trial number, semantic Memory, Word Delayed Recall, Long Term Memory, Orientation, Praxis, Story Recall, Cog Score and WLM.

D. *Classification of diseases*

In this paper the following four classification techniques are used.

1. Naive Bayes classifier
2. Decision tree algorithm J48.
3. Random forest
4. JRIP

1. Naive Bayes classifiers

Naïve Bayesian classifier is a selective classifier which calculates the set of probabilities by counting the frequency and combination of values in a given data set. It assumes that the all variables which contribute towards classification are mutually independent[23].Naïve Bayesian classifier is based on bayes theorem and theorem of total probabilities. Equation 1 is the probability of a document'd' with vector 'x= {x1, x2…xn}' belongs to hypotheses 'h' is given by,

$$P(h1|xi) = \frac{P(xi|h1)P(h1)}{P(xi|h1)P(h1)+P(xi|h2)P9h2)} \quad \ldots (1)$$

P (h1|xi) is the posterior probability and P(h1) is the prior probability associated with hypothesis h1.Equation 2 shows the posterior probability for n different hypotheses, [23].

$$P(h1|xi) = \frac{P(xi|h1)P(h1)}{P(XI)} \quad \ldots (2)$$

$$P(xi) = \sum_{j=1}^{n} P(xi|hj)P(hj) \quad \ldots (3)$$

2. *Decision tree algorithm J48*

J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. In this type of classifier, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple. While building a tree, J48 ignores the missing values i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The basic idea is to divide the data into range based on the attribute values for that item that are found in the training sample. J48 allows classification via either decision trees or rules generated from them [24].

3. Random forest

Random forests are an ensemble learning method for classification that operate by constructing a multitude of decision trees at training time and outputting the class, that is the mode of the classes output by individual trees. In this, the features are randomly selected in each decision split. The correlation between trees is reduced by randomly selecting the features which improves the prediction power and results in higher efficiency. As such the advantages of Random Forest are overcoming the problem of over fitting [25].

4. Rule based classification

In these classifiers the learned model is represented as a set of IF-THEN rules. Rules are good way of representing information. IF-THEN rules is expressed in the form of

IF *condition* THEN *conclusion*

A rule R can be assessed by its coverage and accuracy. Given a tuple, X, from a class-labeled data set, D, let n_{covers} be the number of tuples covered by R; $n_{correct}$ be the number of tuples correctly classified by R; and |D| be the number of tuples in D. The coverage and accuracy of R is defined equation 4 and equation 5.

$$coverage(R) = \frac{\text{ncovers}}{|D|} \quad \ldots (4)$$

$$Accuracy(R) = \frac{\text{ncorrect}}{\text{ncovers}} \quad \ldots (5)$$

That is, a rule's coverage is the percentage of tuples that are covered by the rule [11].

Under this there are various classifiers. Authors have focused on JRIP classifier.

5. JRip Rules Classifiers

JRip (RIPPER) is one of the basic and most popular algorithms. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced error JRip (RIPPER) proceeds by treating all the examples of a particular judgment in the training data as a class, and finding a set of rules that cover all the members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered [26].

V. EXPERIMENTAL RESULTS

The classification techniques are applied on the data set. The parameters like Classification Accuracy, Precession, recall and time taken to build the model are considered. The summary of the data sets are shown in Table 1.

Table1: Summary of datasets

Method	Classification Accuracy	Precession	Recall	Time taken to build the model
Naïve Bayes	100%	1.000	1.000	0.03Sec
J48	98.4%	0.984	0.984	0 Sec
Random forest	100%	1.000	1.000	0.03 Sec
JRIP	100%	1.000	1.000	0.06Sec

The accuracy of the classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

Precession is the percentage of retrieved documents that are in fact relevant to the query. It is defined in equation 6.

$$precesion = \frac{|\{relevant\} \cap \{retrieved\}|}{|\{retrieved\}|} \quad \ldots (6)$$

Recall is the percentage of documents that are relevant to the query were in fact retrieved. It is defined in equation 7.

$$recall = \frac{|\{relevant\} \cap \{retrieved\}|}{|\{relevant\}|} \quad \ldots (7)$$

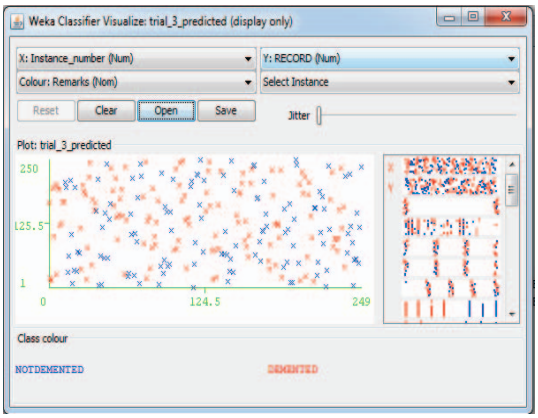

Fig.2: Naïve Bayes classifier error curve

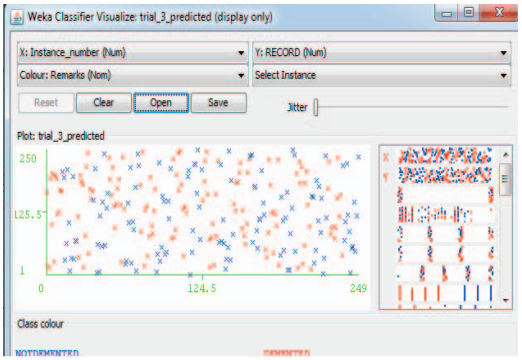

Fig.3: J48 classifier error curve

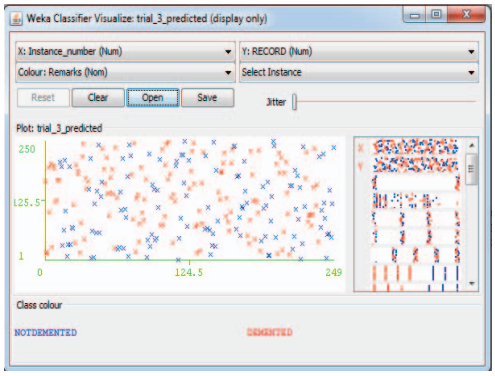

Fig.4: Random forest classifier error curve

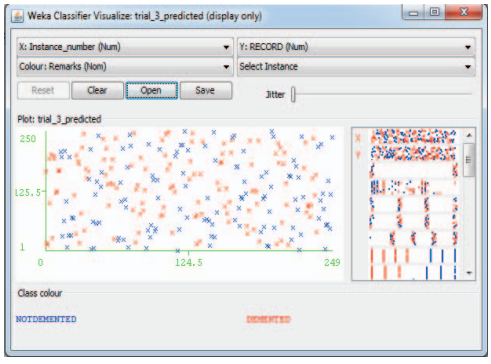

Fig.5: JRIP classifier error curve

VI. CONCLUSION

In this paper different classification techniques of data mining were compared. Classification accuracy, precession, recall and time required for execution of each technique is observed. Performance evaluation of J48, Naïve bayes, random forest and JRIP is tabulated. The authors conclude that Naïve bayes is best of all the four techniques. Fig.6 shows the bar chart of classification accuracy of various machine learning techniques.

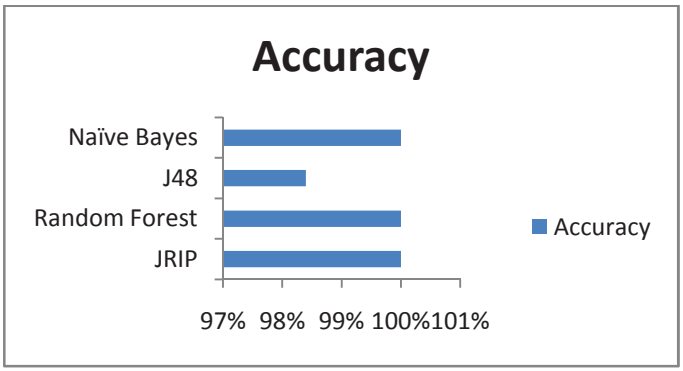

Fig.6: Accuracy of various machine learning techniques

The confusion matrices of different classifications are shown below.

=== Confusion Matrix J48 ===

a	b	<-- classified as
121	2	a = Not Demented
1	126	b = Demented

=== Confusion Matrix JRIP===

a	b	<-- classified as
123	0	a = Not Demented
0	127	b = Demented

=== Confusion Matrix Naïve Bayes===

a	b	<-- classified as
123	0	a = Not Demented
0	127	b = Demented

=== Confusion Matrix Random forest===

a	b	<-- classified as
123	0	a = Not Demented
0	127	b = Demented

Future work includes designing an embedded system based model to diagnose the patient.

ACKNOWLEDGEMENT

The authors are thankful to Dr. Sandhya Joshi, Associate professor, VTU, Gulbarga for her support and guidance. Dr. Murali Krishna, Earlier Scientist Research Fellow, Wellcome DBT Allianz, CSI Holdsworth Memorial Mission Hospital, Mysore, Dr. L Basavaraj, Principal, ATME, Mysore and to the research colleagues who supported with the data in respect of the Alzheimer`s disease.

REFERENCE

[1] David P Salnom and Mark W.Bondi "Neuropsychological Assessment of Dementia" Access NIH public, PubMed central, US national library of Medicine National Institutes of Health, May 2010.

[2] http://www.capitalfm.co.ke/lifestyle/2013/12/06/44-million-now-suffer-from-dementia-worldwide

[3] Viswanathan A, Rocca WA, Tzourio C. Vascular risk factors and dementia: How to move forward? Neurology 72:Pp368–74, 2009

[4] http://www.todayszaman.com/news-293073-38-million-people-suffering-from-alzheimers-disease-worldwide.html

[5] Bhagya Shree S. R, Dr. H. S. Sheshadri "An Approach in the Diagnosis of Alzheimer Disease - A Survey" International Journal of Engineering Trends and Technology (IJETT) – Volume 7 Number 1- Jan 2014 ISSN: 2231

[6] Michael saling, Henry Brodaty, Dr. Mark Yates, Dr. Sam Scherer, Professor Kaarin Anstey, "Early Diagnosis of Dementia", 2007.

[7]http://alzheimers.emedtv.com/dementia/dementia-risk-factors.html

[8] Thies W, Bleiler L, "2013Alzheimer's Facts and figures" Alzheimer's Dement (Journal of Alzheimer's association), Elsevier Inc. Mar-2013.

[9] Bhagya Shree S R, Dr. H. S. Sheshadri and Dr. Sandhya Joshi "A Review on the Method of Diagnosing Alzheimer's Disease using Data Mining" International Journal of Engineering Research & Technology (IJERT),Vol. 3 Issue 3,Pp 2417March - 2014

[10]Ana Luisa Sosa, Emiliano Albanese, Martin Prince, Daisy Acosta, Cleusa P Ferri, Mariella Guerra, Yueqin Huang, K S Jacob, Juan Llibre de Rodriguez, Aquiles Salas, Fang Yang, CiroGaona, A T Joteeshwaran, Guillermina Rodriguez, Gabriela Rojas de la Torre, Joseph D Williams and Robert Stewart "Population normative data for the 10/66 Dementia Research Group cognitive test battery from Latin America, India and China: across-sectional survey" BMC Neurology, Vol9, pp 1-11, Aug 2009

[11] Jiawei Han, Micheline Kamber, JianPei "*Data Mining: Concepts and Techniques*" published by Elsevier, Third edition, 2012

[12] G K Gupta "Introduction to data mining with case studies" Eastern economy edition, second edition.

[13] Jyoti Soni, Ujma Ansari, Dipesh Sharma and Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011

[14] Ruijuan Hu "Medical Data Mining Based on Decision Tree Algorithm", Computer and Information Science Vol. 4, No. 5; September 2011

[15] Amir Fallahi and Shahram Jafari "An Expert System for Detection of Breast Cancer Using Data Preprocessing and Bayesian Network", International Journal of Advanced Science and Technology Vol. 34, September, 2011

[16] Abhishek Taneja "Heart Disease Prediction System" Oriental journal of computer science & technology ISSN: 0974-6471 December 2013, Vol. 6, No. (4)

[17] Duarte Ferreira, Abílio Oliveira and Alberto Freitas "Applying data mining techniques to improve diagnosis in neonatal jaundice" ,Bio medical central, 2012

[18] P. Rajeswari and G.Sophia Reena "Analysis of Liver Disorder Using Data mining Algorithm" Global journal of computer science and technology Vol. 10 Issue 14 (Ver. 1.0) November 2010

[19] Tarigoppula V.S Sriram, M. Venkateswara Rao, G V Satya Narayana, DSVGK Kaladhar and T Pandu Ranga Vital "Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms" International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 3, September 2013

[20] Sandhya Joshi, P. Deepa Shenoy, Vibhudendra Simha G.G., Venugopal K. R and L. M. Patnaik "Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques" IACSIT International Journal of Engineering and Technology, Vol.2, No.4, ISSN: 1793-8236, August 2010

[21] Javier Escudero, Emmanuel I Feachor, and Stephen Pearson, "Machine Learning-Based Method for Personalized and Cost-Effective Detection of Alzheimer's Disease" IEEE transactions on biomedical engineering, vol. 60, no. 1, January 2013

[22] Plamena Andreeva, Maya Dimitrova and PetiaRadeva "data mining learning models and algorithms for medical applications"

[23] K.P.Soman, Shyam Diwakar, V. Ajay "*Insight into data mining theory and practice",* New Delhi PHI learning Pvt. Ltd

[24] Tina R. Patil and Mrs. S. S. Sherekar "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification" International Journal of Computer Science and Applications Vol. 6, No.2, ISSN: 0974-1011, Pp256, April 2013

[25] Jehad Ali Rehanullah Khan, Nasir Ahmad, Imran Maqsood "Random Forests and Decision Trees" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, ISSN: 1694-0814, September 2012

[26] Anil Rajput, Ramesh Prasad Aharwal, Meghna Dubey S.P. Saxena and ManmohanRaghuvanshi"J48 and JRIP Rules for E-Governance Data" International Journal of Computer Science and Security (IJCSS), Volume (5): Issue (2), Pp 201, 2011.